孙亚威

ywsun@ir.hit.edu.cn · (+86) 133-820-55988 · Yawei Sun

教育背景

哈尔滨工业大学, 哈尔滨 2019 – 至今

在读硕士研究生 计算机科学与技术, 在社会计算与信息检索实验室进行自然语言文本生成技术研究, 师从秦兵教授, 预计 2021 年 7 月毕业

哈尔滨工业大学, 哈尔滨 2015 – 2019

学士 计算机科学与技术

实习/项目/科研经历

哈尔滨深智科技有限公司 2018 年 8 月 – 2018 年 12 月

实习 经理: 史华兴

自然语言处理算法工程师

- 基于论文 Multiway Attention Networks for Modeling Sentence Pairs 实现的检索型自动评论系统, github 项目地址: https://github.com/syw1996/Retrieval-Automatic-Comment-System
- 基于论文 Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative ShortText Conversation 实现基于关键词的对话机器人, github 项目地址: https://github.com/syw1996/Seq2BF-pytorch
- 基于论文 Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory，将其 TensorFlow 版代码迁移至 Pytorch

深圳鹏城实验室 2019 年 2 月 – 2019 年 5 月

实习 参与基于 Rotowire 数据集的表格到文本生成的研究

- 基于篇章级表格到文本生成以及无监督文本风格迁移的工作，在导师冯骁骋指导下撰写论文 Learning to Select Bi-Aspect Information for Document-Scale Text Content Manipulation，该论文已被 CCF A 类会议 AAAI 2020 收录，第一作者是导师冯骁骋，本人是第二作者，github 项目地址: https://github.com/syw1996/SCIR-TG-Data2text-Bi-Aspect

腾讯小说生成项目 2018 年 9 月 – 2019 年 2 月

Python OpenNMT 实验室与腾讯合作项目

- 设计一个端到端的小说生成模型，基于输入的外貌关键词生成小说人物描写片段
- 引入强化学习中的策略梯度机制，以生产文本的 BLEU 值作为强化学习的奖励
- 尝试引入无监督文本风格迁移机制，能够极大提高人物描写片段风格的多样性
- 该项目已经成功申请腾讯方面专利《一种基于深度学习技术的段落级文本外貌描写自动生成发明》，并被应用在实验室的文本生成平台

基于 GPT-2 的维基百科表格到文本生成 2020 年 2 月 – 2020 年 6 月

Python Transformers 和实验室龚恒师兄以共一作身份投稿 CCF B 类会议 Coling 2020，目前已被收录

- 尝试利用 GPT-2 预训练语言生成模型在 low resource Wiki 百科数据集上进行表格到文本生成任务，缓解传统表格到文本生成任务对大量领域内数据的依赖，在 Human、Books 以及 Songs 三个领域数据集上均取得 SOTA 结果
- 利用自然语言模板对半结构化的 Wiki 表格进行线性序列化处理，使其符合 GPT-2 模型输入规范
- 在微调阶段引入表格记录属性分类任务，提高 GPT-2 模型对表格各个不同属性记录的建模能力
- 在微调阶段引入内容匹配任务，利用 Optimal Transport 机制计算表格与生成文本的 Wasserstein 距离，提高生成文本内容真实度并减少生成信息冗余度

基于数字推理的表格到文本生成 2019 年 9 月 – 2019 年 12 月

Python Pytorch 参与部分实验设计实施，目前 ACL 2020 被拒稿，以署名身份转投期刊

- 在 Rotowire 数据集上进行篇章级结构化表格到文本生成任务，在部分关键事实类评价指标上取得 SOTA 结果
- 对于报道文本中部分比赛记录数字需要根据对应表格中相关数字进行推理才能得到的问题，提出了在解码阶段基于输入比赛表格生成数学计算等式，进而推理生成出表格中没有的数字记录
- 为了让模型的结构化表格编码器从更好的起点启动训练，设计了一种预训练机制，输入 MASK 掉部分数字记录的表格，让模型依据表格中其他的数字信息复原 MASK 掉的数字信息，在正式训练阶段利用预训练得到编码器的参数初始化

IT 技能

- 编程语言: Python
- 深度学习库: Pytorch > TensorFlow
- Web 开发: Flask

♡ 获奖情况

全国二等奖, 全国大学生信息安全竞赛 2018 年 7 月
创新组特等奖, 第四届全国青年人工智能创新创业 2018 年 12 月
一等奖学金, 哈尔滨工业大学研究生专项奖学金 2019 年 9 月
联想奖学金, 2019-2020 年度联想奖学金 2020 年 3 月

其他

- GitHub: https://github.com/syw1996
- 谷歌学术: https://scholar.google.com.hk/citations?hl=zh-CN&user=HcNGFksAAAAJ
- 语言: 英语 - 熟练 (CET-6 550，TOEFL 84)